```
Assignment no. -02
Aim-1. Creation of Dataset using Microsoft Excel.
2. Identification and Handling of Null Values
3. Identification and Handling of Outliers
4. Data Transformation for the purpose of :
a. To change the scale for better understanding
b. To decrease the skewness and convert distribution into normal
distribution
```

In [1]:
```
import pandas as pd
import seaborn as sns
import numpy as py
```

In [4]:
```
df= pd.read_csv("C:/Users/Welcome/Music/Book1.csv")
```

In [5]:
```
df
```

Out[5]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60.0	63.0	76.0	95.0	2021	3	female
1	75.0	70.0	64.0	85.0	2020	3	male
2	74.0	50.0	55.0	91.0	2020	3	male
3	68.0	76.0	78.0	97.0	2020	3	female
4	NaN	67.0	71.0	93.0	2020	3	male
5	70.0	64.0	80.0	98.0	2018	3	female
6	61.0	78.0	92.0	94.0	2021	3	male
7	61.0	74.0	78.0	NaN	2021	2	male
8	64.0	76.0	79.0	76.0	2019	2	male
9	65.0	95.0	75.0	90.0	2020	3	female
10	66.0	76.0	NaN	100.0	2019	1	male
11	84.0	67.0	71.0	92.0	2020	3	male
12	69.0	NaN	70.0	86.0	2021	3	female
13	74.0	65.0	65.0	80.0	2021	2	male
14	74.0	63.0	72.0	96.0	2018	3	male
15	76.0	64.0	80.0	96.0	2020	3	male
16	60.0	64.0	54.0	91.0	2021	3	female
17	77.0	70.0	72.0	99.0	2020	3	male
18	67.0	95.0	NaN	87.0	2018	3	female
19	71.0	53.0	78.0	75.0	2018	2	female
20	58.0	65.0	56.0	NaN	2019	3	female
21	68.0	63.0	62.0	94.0	2021	3	male
22	77.0	63.0	68.0	97.0	2021	3	female
23	80.0	NaN	86.0	85.0	2018	3	female
24	84.0	63.0	67.0	83.0	2018	1	male
25	68.0	67.0	73.0	88.0	2019	3	female
26	76.0	64.0	68.0	96.0	2021	3	female
27	92.0	96.0	61.0	83.0	2018	2	male
28	60.0	68.0	59.0	93.0	2020	3	male

In [6]:
```
df.isnull()
```

Out[6]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False
4	True	False	False	False	False	False	False
5	False	False	False	False	False	False	False
6	False	False	False	False	False	False	False
7	False	False	False	True	False	False	False
8	False	False	False	False	False	False	False
9	False	False	False	False	False	False	False
10	False	False	True	False	False	False	False
11	False	False	False	False	False	False	False
12	False	True	False	False	False	False	False
13	False	False	False	False	False	False	False
14	False	False	False	False	False	False	False
15	False	False	False	False	False	False	False
16	False	False	False	False	False	False	False
17	False	False	False	False	False	False	False
18	False	False	True	False	False	False	False
19	False	False	False	False	False	False	False
20	False	False	False	True	False	False	False
21	False	False	False	False	False	False	False
22	False	False	False	False	False	False	False
23	False	True	False	False	False	False	False
24	False	False	False	False	False	False	False
25	False	False	False	False	False	False	False
26	False	False	False	False	False	False	False
27	False	False	False	False	False	False	False
28	False	False	False	False	False	False	False

In [7]:
```
series = pd.isnull(df["math score"])
df[series]
```

Out[7]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
4	NaN	67.0	71.0	93.0	2020	3	male

In [8]:
```
df.notnull()
```

Out[8]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	True	True	True	True	True	True	True
1	True	True	True	True	True	True	True
2	True	True	True	True	True	True	True
3	True	True	True	True	True	True	True
4	False	True	True	True	True	True	True
5	True	True	True	True	True	True	True
6	True	True	True	True	True	True	True
7	True	True	True	False	True	True	True
8	True	True	True	True	True	True	True
9	True	True	True	True	True	True	True
10	True	True	False	True	True	True	True
11	True	True	True	True	True	True	True
12	True	False	True	True	True	True	True
13	True	True	True	True	True	True	True
14	True	True	True	True	True	True	True
15	True	True	True	True	True	True	True
16	True	True	True	True	True	True	True
17	True	True	True	True	True	True	True
18	True	True	False	True	True	True	True
19	True	True	True	True	True	True	True
20	True	True	True	False	True	True	True
21	True	True	True	True	True	True	True
22	True	True	True	True	True	True	True
23	True	False	True	True	True	True	True
24	True	True	True	True	True	True	True
25	True	True	True	True	True	True	True
26	True	True	True	True	True	True	True
27	True	True	True	True	True	True	True
28	True	True	True	True	True	True	True

In [9]:
```
series1 = pd.notnull(df["math score"])
df[series1]
```

Out[9]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60.0	63.0	76.0	95.0	2021	3	female
1	75.0	70.0	64.0	85.0	2020	3	male
2	74.0	50.0	55.0	91.0	2020	3	male
3	68.0	76.0	78.0	97.0	2020	3	female
5	70.0	64.0	80.0	98.0	2018	3	female
6	61.0	78.0	92.0	94.0	2021	3	male
7	61.0	74.0	78.0	NaN	2021	2	male
8	64.0	76.0	79.0	76.0	2019	2	male
9	65.0	95.0	75.0	90.0	2020	3	female
10	66.0	76.0	NaN	100.0	2019	1	male
11	84.0	67.0	71.0	92.0	2020	3	male
12	69.0	NaN	70.0	86.0	2021	3	female
13	74.0	65.0	65.0	80.0	2021	2	male
14	74.0	63.0	72.0	96.0	2018	3	male
15	76.0	64.0	80.0	96.0	2020	3	male
16	60.0	64.0	54.0	91.0	2021	3	female
17	77.0	70.0	72.0	99.0	2020	3	male
18	67.0	95.0	NaN	87.0	2018	3	female
19	71.0	53.0	78.0	75.0	2018	2	female
20	58.0	65.0	56.0	NaN	2019	3	female
21	68.0	63.0	62.0	94.0	2021	3	male
22	77.0	63.0	68.0	97.0	2021	3	female
23	80.0	NaN	86.0	85.0	2018	3	female
24	84.0	63.0	67.0	83.0	2018	1	male
25	68.0	67.0	73.0	88.0	2019	3	female
26	76.0	64.0	68.0	96.0	2021	3	female
27	92.0	96.0	61.0	83.0	2018	2	male
28	60.0	68.0	59.0	93.0	2020	3	male

In [10]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['gender'] = le.fit_transform(df['gender'])
newdf = df
df
```

Out[10]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60.0	63.0	76.0	95.0	2021	3	0
1	75.0	70.0	64.0	85.0	2020	3	1
2	74.0	50.0	55.0	91.0	2020	3	1
3	68.0	76.0	78.0	97.0	2020	3	0
4	NaN	67.0	71.0	93.0	2020	3	1
5	70.0	64.0	80.0	98.0	2018	3	0
6	61.0	78.0	92.0	94.0	2021	3	1
7	61.0	74.0	78.0	NaN	2021	2	1
8	64.0	76.0	79.0	76.0	2019	2	1
9	65.0	95.0	75.0	90.0	2020	3	0
10	66.0	76.0	NaN	100.0	2019	1	1

In [11]:
```
missing_values = ["Na","na"]
df = pd.read_csv("C:/Users/Welcome/Music/Book1.csv", na_values = missing_va
df
```

Out[11]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60.0	63.0	76.0	95.0	2021	3	female
1	75.0	70.0	64.0	85.0	2020	3	male
2	74.0	50.0	55.0	91.0	2020	3	male
3	68.0	76.0	78.0	97.0	2020	3	female
4	NaN	67.0	71.0	93.0	2020	3	male
5	70.0	64.0	80.0	98.0	2018	3	female
6	61.0	78.0	92.0	94.0	2021	3	male
7	61.0	74.0	78.0	NaN	2021	2	male
8	64.0	76.0	79.0	76.0	2019	2	male
9	65.0	95.0	75.0	90.0	2020	3	female
10	66.0	76.0	NaN	100.0	2019	1	male
11	84.0	67.0	71.0	92.0	2020	3	male
12	69.0	NaN	70.0	86.0	2021	3	female
13	74.0	65.0	65.0	80.0	2021	2	male
14	74.0	63.0	72.0	96.0	2018	3	male
15	76.0	64.0	80.0	96.0	2020	3	male
16	60.0	64.0	54.0	91.0	2021	3	female
17	77.0	70.0	72.0	99.0	2020	3	male
18	67.0	95.0	NaN	87.0	2018	3	female
19	71.0	53.0	78.0	75.0	2018	2	female
20	58.0	65.0	56.0	NaN	2019	3	female
21	68.0	63.0	62.0	94.0	2021	3	male
22	77.0	63.0	68.0	97.0	2021	3	female
23	80.0	NaN	86.0	85.0	2018	3	female
24	84.0	63.0	67.0	83.0	2018	1	male
25	68.0	67.0	73.0	88.0	2019	3	female
26	76.0	64.0	68.0	96.0	2021	3	female
27	92.0	96.0	61.0	83.0	2018	2	male
28	60.0	68.0	59.0	93.0	2020	3	male

In [12]:
```
ndf = df
ndf.fillna(0)
```

Out[12]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60.0	63.0	76.0	95.0	2021	3	female
1	75.0	70.0	64.0	85.0	2020	3	male
2	74.0	50.0	55.0	91.0	2020	3	male
3	68.0	76.0	78.0	97.0	2020	3	female
4	0.0	67.0	71.0	93.0	2020	3	male
5	70.0	64.0	80.0	98.0	2018	3	female
6	61.0	78.0	92.0	94.0	2021	3	male
7	61.0	74.0	78.0	0.0	2021	2	male
8	64.0	76.0	79.0	76.0	2019	2	male
9	65.0	95.0	75.0	90.0	2020	3	female
10	66.0	76.0	0.0	100.0	2019	1	male
11	84.0	67.0	71.0	92.0	2020	3	male
12	69.0	0.0	70.0	86.0	2021	3	female
13	74.0	65.0	65.0	80.0	2021	2	male
14	74.0	63.0	72.0	96.0	2018	3	male
15	76.0	64.0	80.0	96.0	2020	3	male
16	60.0	64.0	54.0	91.0	2021	3	female
17	77.0	70.0	72.0	99.0	2020	3	male
18	67.0	95.0	0.0	87.0	2018	3	female
19	71.0	53.0	78.0	75.0	2018	2	female
20	58.0	65.0	56.0	0.0	2019	3	female
21	68.0	63.0	62.0	94.0	2021	3	male
22	77.0	63.0	68.0	97.0	2021	3	female
23	80.0	0.0	86.0	85.0	2018	3	female
24	84.0	63.0	67.0	83.0	2018	1	male
25	68.0	67.0	73.0	88.0	2019	3	female
26	76.0	64.0	68.0	96.0	2021	3	female
27	92.0	96.0	61.0	83.0	2018	2	male
28	60.0	68.0	59.0	93.0	2020	3	male

In [13]:
```
m_v=df['math score'].mean()
df['math score'].fillna(value = m_v, inplace = True)
df
```

Out[13]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60.000000	63.0	76.0	95.0	2021	3	female
1	75.000000	70.0	64.0	85.0	2020	3	male
2	74.000000	50.0	55.0	91.0	2020	3	male
3	68.000000	76.0	78.0	97.0	2020	3	female
4	70.678571	67.0	71.0	93.0	2020	3	male
5	70.000000	64.0	80.0	98.0	2018	3	female
6	61.000000	78.0	92.0	94.0	2021	3	male
7	61.000000	74.0	78.0	NaN	2021	2	male
8	64.000000	76.0	79.0	76.0	2019	2	male
9	65.000000	95.0	75.0	90.0	2020	3	female
10	66.000000	76.0	NaN	100.0	2019	1	male
11	84.000000	67.0	71.0	92.0	2020	3	male
12	69.000000	NaN	70.0	86.0	2021	3	female
13	74.000000	65.0	65.0	80.0	2021	2	male
14	74.000000	63.0	72.0	96.0	2018	3	male
15	76.000000	64.0	80.0	96.0	2020	3	male
16	60.000000	64.0	54.0	91.0	2021	3	female
17	77.000000	70.0	72.0	99.0	2020	3	male
18	67.000000	95.0	NaN	87.0	2018	3	female
19	71.000000	53.0	78.0	75.0	2018	2	female
20	58.000000	65.0	56.0	NaN	2019	3	female
21	68.000000	63.0	62.0	94.0	2021	3	male
22	77.000000	63.0	68.0	97.0	2021	3	female
23	80.000000	NaN	86.0	85.0	2018	3	female
24	84.000000	63.0	67.0	83.0	2018	1	male
25	68.000000	67.0	73.0	88.0	2019	3	female
26	76.000000	64.0	68.0	96.0	2021	3	female
27	92.000000	96.0	61.0	83.0	2018	2	male
28	60.000000	68.0	59.0	93.0	2020	3	male

In [14]:
```
ndf.replace(to_replace = py.nan, value = -99)
```

Out[14]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60.000000	63.0	76.0	95.0	2021	3	female
1	75.000000	70.0	64.0	85.0	2020	3	male
2	74.000000	50.0	55.0	91.0	2020	3	male
3	68.000000	76.0	78.0	97.0	2020	3	female
4	70.678571	67.0	71.0	93.0	2020	3	male
5	70.000000	64.0	80.0	98.0	2018	3	female
6	61.000000	78.0	92.0	94.0	2021	3	male
7	61.000000	74.0	78.0	-99.0	2021	2	male
8	64.000000	76.0	79.0	76.0	2019	2	male
9	65.000000	95.0	75.0	90.0	2020	3	female
10	66.000000	76.0	-99.0	100.0	2019	1	male
11	84.000000	67.0	71.0	92.0	2020	3	male
12	69.000000	-99.0	70.0	86.0	2021	3	female
13	74.000000	65.0	65.0	80.0	2021	2	male
14	74.000000	63.0	72.0	96.0	2018	3	male
15	76.000000	64.0	80.0	96.0	2020	3	male
16	60.000000	64.0	54.0	91.0	2021	3	female
17	77.000000	70.0	72.0	99.0	2020	3	male
18	67.000000	95.0	-99.0	87.0	2018	3	female
19	71.000000	53.0	78.0	75.0	2018	2	female
20	58.000000	65.0	56.0	-99.0	2019	3	female
21	68.000000	63.0	62.0	94.0	2021	3	male
22	77.000000	63.0	68.0	97.0	2021	3	female
23	80.000000	-99.0	86.0	85.0	2018	3	female
24	84.000000	63.0	67.0	83.0	2018	1	male
25	68.000000	67.0	73.0	88.0	2019	3	female
26	76.000000	64.0	68.0	96.0	2021	3	female
27	92.000000	96.0	61.0	83.0	2018	2	male
28	60.000000	68.0	59.0	93.0	2020	3	male

In [15]:
```
ndf.dropna()
```

Out[15]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60.000000	63.0	76.0	95.0	2021	3	female
1	75.000000	70.0	64.0	85.0	2020	3	male
2	74.000000	50.0	55.0	91.0	2020	3	male
3	68.000000	76.0	78.0	97.0	2020	3	female
4	70.678571	67.0	71.0	93.0	2020	3	male
5	70.000000	64.0	80.0	98.0	2018	3	female
6	61.000000	78.0	92.0	94.0	2021	3	male
8	64.000000	76.0	79.0	76.0	2019	2	male
9	65.000000	95.0	75.0	90.0	2020	3	female
11	84.000000	67.0	71.0	92.0	2020	3	male
13	74.000000	65.0	65.0	80.0	2021	2	male
14	74.000000	63.0	72.0	96.0	2018	3	male
15	76.000000	64.0	80.0	96.0	2020	3	male
16	60.000000	64.0	54.0	91.0	2021	3	female
17	77.000000	70.0	72.0	99.0	2020	3	male
19	71.000000	53.0	78.0	75.0	2018	2	female
21	68.000000	63.0	62.0	94.0	2021	3	male
22	77.000000	63.0	68.0	97.0	2021	3	female
24	84.000000	63.0	67.0	83.0	2018	1	male
25	68.000000	67.0	73.0	88.0	2019	3	female
26	76.000000	64.0	68.0	96.0	2021	3	female
27	92.000000	96.0	61.0	83.0	2018	2	male
28	60.000000	68.0	59.0	93.0	2020	3	male

In [16]:
```
ndf.dropna(how = 'all')
```

Out[16]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60.000000	63.0	76.0	95.0	2021	3	female
1	75.000000	70.0	64.0	85.0	2020	3	male
2	74.000000	50.0	55.0	91.0	2020	3	male
3	68.000000	76.0	78.0	97.0	2020	3	female
4	70.678571	67.0	71.0	93.0	2020	3	male
5	70.000000	64.0	80.0	98.0	2018	3	female
6	61.000000	78.0	92.0	94.0	2021	3	male
7	61.000000	74.0	78.0	NaN	2021	2	male
8	64.000000	76.0	79.0	76.0	2019	2	male
9	65.000000	95.0	75.0	90.0	2020	3	female
10	66.000000	76.0	NaN	100.0	2019	1	male
11	84.000000	67.0	71.0	92.0	2020	3	male
12	69.000000	NaN	70.0	86.0	2021	3	female
13	74.000000	65.0	65.0	80.0	2021	2	male
14	74.000000	63.0	72.0	96.0	2018	3	male
15	76.000000	64.0	80.0	96.0	2020	3	male
16	60.000000	64.0	54.0	91.0	2021	3	female
17	77.000000	70.0	72.0	99.0	2020	3	male
18	67.000000	95.0	NaN	87.0	2018	3	female
19	71.000000	53.0	78.0	75.0	2018	2	female
20	58.000000	65.0	56.0	NaN	2019	3	female
21	68.000000	63.0	62.0	94.0	2021	3	male
22	77.000000	63.0	68.0	97.0	2021	3	female
23	80.000000	NaN	86.0	85.0	2018	3	female
24	84.000000	63.0	67.0	83.0	2018	1	male
25	68.000000	67.0	73.0	88.0	2019	3	female
26	76.000000	64.0	68.0	96.0	2021	3	female
27	92.000000	96.0	61.0	83.0	2018	2	male
28	60.000000	68.0	59.0	93.0	2020	3	male

In [17]:
```
ndf.dropna(axis = 1)
```

Out[17]:

	math score	club join year	placement offer count	gender
0	60.000000	2021	3	female
1	75.000000	2020	3	male
2	74.000000	2020	3	male
3	68.000000	2020	3	female
4	70.678571	2020	3	male
5	70.000000	2018	3	female
6	61.000000	2021	3	male
7	61.000000	2021	2	male
8	64.000000	2019	2	male
9	65.000000	2020	3	female
10	66.000000	2019	1	male
11	84.000000	2020	3	male
12	69.000000	2021	3	female
13	74.000000	2021	2	male
14	74.000000	2018	3	male
15	76.000000	2020	3	male
16	60.000000	2021	3	female
17	77.000000	2020	3	male
18	67.000000	2018	3	female
19	71.000000	2018	2	female
20	58.000000	2019	3	female
21	68.000000	2021	3	male
22	77.000000	2021	3	female
23	80.000000	2018	3	female
24	84.000000	2018	1	male
25	68.000000	2019	3	female
26	76.000000	2021	3	female
27	92.000000	2018	2	male
28	60.000000	2020	3	male

In [18]:
```
new_data = ndf.dropna(axis = 0, how='any')
new_data
```

Out[18]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60.000000	63.0	76.0	95.0	2021	3	female
1	75.000000	70.0	64.0	85.0	2020	3	male
2	74.000000	50.0	55.0	91.0	2020	3	male
3	68.000000	76.0	78.0	97.0	2020	3	female
4	70.678571	67.0	71.0	93.0	2020	3	male
5	70.000000	64.0	80.0	98.0	2018	3	female
6	61.000000	78.0	92.0	94.0	2021	3	male
8	64.000000	76.0	79.0	76.0	2019	2	male
9	65.000000	95.0	75.0	90.0	2020	3	female
11	84.000000	67.0	71.0	92.0	2020	3	male
13	74.000000	65.0	65.0	80.0	2021	2	male
14	74.000000	63.0	72.0	96.0	2018	3	male
15	76.000000	64.0	80.0	96.0	2020	3	male
16	60.000000	64.0	54.0	91.0	2021	3	female
17	77.000000	70.0	72.0	99.0	2020	3	male
19	71.000000	53.0	78.0	75.0	2018	2	female
21	68.000000	63.0	62.0	94.0	2021	3	male
22	77.000000	63.0	68.0	97.0	2021	3	female
24	84.000000	63.0	67.0	83.0	2018	1	male
25	68.000000	67.0	73.0	88.0	2019	3	female
26	76.000000	64.0	68.0	96.0	2021	3	female
27	92.000000	96.0	61.0	83.0	2018	2	male
28	60.000000	68.0	59.0	93.0	2020	3	male

In [19]:
```
import matplotlib.pyplot as plt
```

In [20]:
```
df1= pd.read_csv("C:/Users/Welcome/Music/Book3.csv")
df1
```

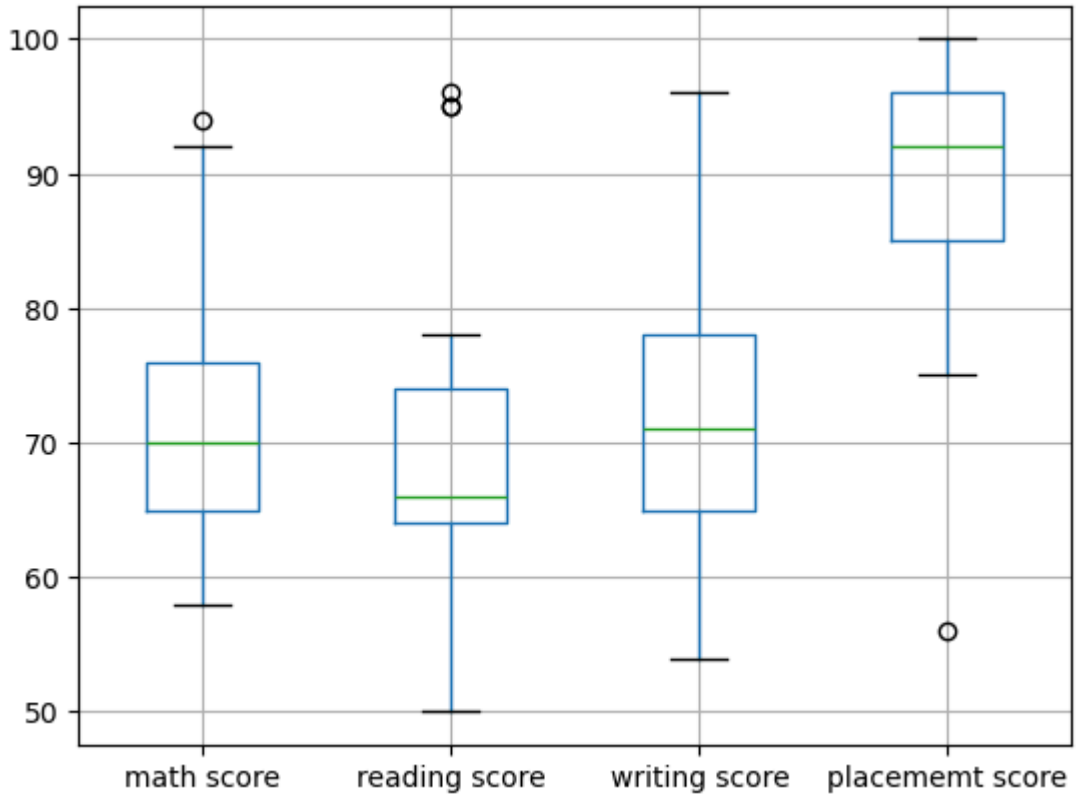
Out[20]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60	63	76	95	2021	3	female
1	75	70	64	85	2020	3	male
2	74	50	55	91	2020	3	male
3	68	76	78	97	2020	3	female
4	94	67	71	93	2020	3	male
5	70	64	80	98	2018	3	female
6	61	78	92	94	2021	3	male
7	61	74	78	80	2021	2	male
8	64	76	79	76	2019	2	male
9	65	95	75	90	2020	3	female
10	66	76	67	100	2019	1	male
11	84	67	71	92	2020	3	male
12	69	66	70	56	2021	3	female
13	74	65	65	80	2021	2	male
14	74	63	72	96	2018	3	male
15	76	64	80	96	2020	3	male
16	60	64	54	91	2021	3	female
17	77	70	72	99	2020	3	male
18	67	95	64	87	2018	3	female
19	71	65	78	75	2018	2	female
20	58	65	96	92	2019	3	female
21	68	63	62	94	2021	3	male
22	77	63	68	97	2021	3	female
23	80	64	86	85	2018	3	female
24	84	63	67	83	2018	1	male
25	68	67	73	88	2019	3	female
26	76	64	68	96	2021	3	female
27	92	96	61	83	2018	2	male
28	60	68	59	93	2020	3	male

In [ ]:

In [21]:
```
col = ['math score', 'reading score', 'writing score', 'placememt score']
df1.boxplot(col)
```

Out[21]: <Axes: >



In [22]:
```
print(py.where(df1['math score']>90))
print(py.where(df1['reading score']<25))
print(py.where(df1['writing score']<30))
```

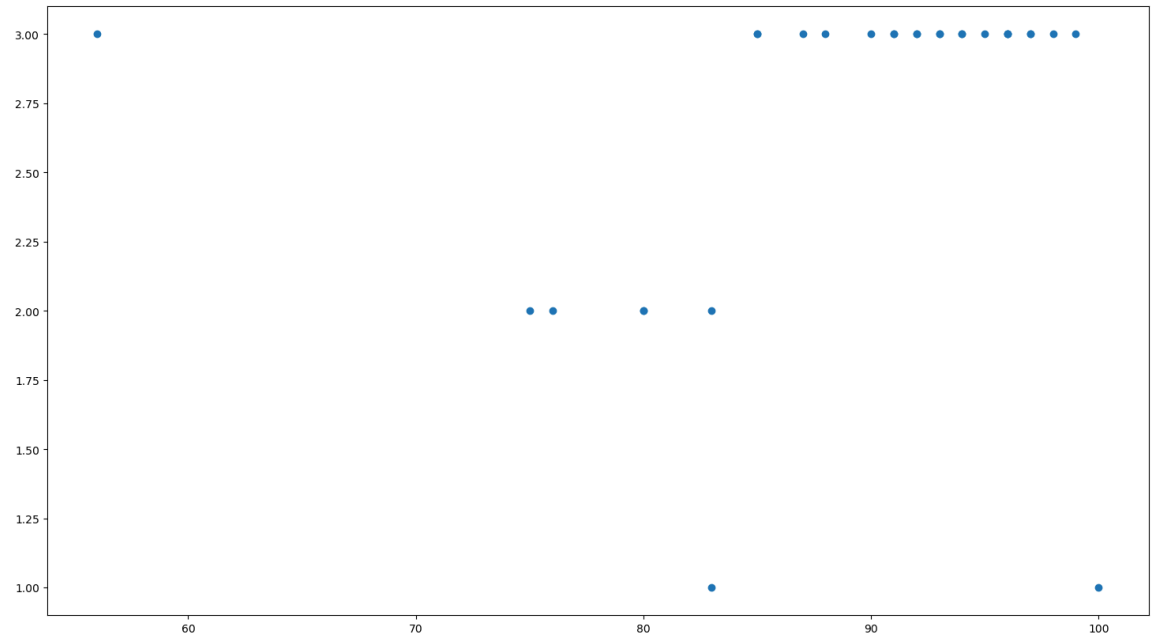
```
(array([ 4, 27], dtype=int64),)
(array([], dtype=int64),)
(array([], dtype=int64),)
```

In [ ]:

In [23]:
```
pip install matplotlib
```

```
Requirement already satisfied: matplotlib in c:\users\welcome\anaconda3\li
b\site-packages (3.8.0)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\welcome\anacon
da3\lib\site-packages (from matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in c:\users\welcome\anaconda3
\lib\site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\welcome\anaco
nda3\lib\site-packages (from matplotlib) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\welcome\anaco
nda3\lib\site-packages (from matplotlib) (1.4.4)
Requirement already satisfied: numpy<2,>=1.21 in c:\users\welcome\anaconda
3\lib\site-packages (from matplotlib) (1.26.4)
Requirement already satisfied: packaging>=20.0 in c:\users\welcome\anacond
a3\lib\site-packages (from matplotlib) (23.1)
Requirement already satisfied: pillow>=6.2.0 in c:\users\welcome\anaconda3
\lib\site-packages (from matplotlib) (10.2.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\welcome\anacon
da3\lib\site-packages (from matplotlib) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\welcome\an
aconda3\lib\site-packages (from matplotlib) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\welcome\anaconda3\lib
\site-packages (from python-dateutil>=2.7->matplotlib) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [24]:
```
fig, ax= plt.subplots(figsize = (18, 10))
ax.scatter(df1['placememt score'], df1['placement offer count'])
plt.show()
ax.set_xlabel('(Proportion non-retail business acres)/(town)')
ax.set_ylabel('(Full-value property-tax rate)/($10,000)')
```



Out[24]:
```
Text(4.444444444444452, 0.5, '(Full-value property-tax rate)/($10,000)')
```

In [25]:
```
print(py.where((df1['placememt score']<50) & (df1['placement offer count']>1
print(py.where((df1['placememt score']>85) & (df1['placement offer count']<1
```

```
(array([], dtype=int64),)
(array([10], dtype=int64),)
```

In [26]:
```
from scipy import stats
```

In [27]:
```
z = py.abs(stats.zscore(df1['math score']))
```

In [28]:
```
print(z)
```

```
0     1.252553
1     0.383665
2     0.274584
3     0.379903
4     2.456207
5     0.161741
6     1.143471
7     1.143471
8     0.816228
9     0.707147
10    0.598066
11    1.365395
12    0.270822
13    0.274584
14    0.274584
15    0.492746
16    1.252553
17    0.601827
18    0.488984
19    0.052660
20    1.470715
21    0.379903
22    0.601827
23    0.929071
24    1.365395
25    0.379903
26    0.492746
27    2.238044
28    1.252553
Name: math score, dtype: float64
```

In [29]:
```
threshold = 0.18
```

In [30]:
```
sample_outliers = py.where(z<threshold)
sample_outliers
```

Out[30]:
```
(array([ 5, 19], dtype=int64),)
```

In [31]:
```
sorted_rscore = sorted(df1['reading score'])
```

In [32]:
```
sorted_rscore
```

Out[32]:
```
[50,
 63,
 63,
 63,
 63,
 63,
 64,
 64,
 64,
 64,
 64,
 65,
 65,
 65,
 66,
 67,
 67,
 67,
 68,
 70,
 70,
 74,
 76,
 76,
 76,
 78,
 95,
 95,
 96]
```

In [33]:
```
q1 = py.percentile(sorted_rscore, 25)
q3 = py.percentile(sorted_rscore, 75)
print(q1, q3)
```

```
64.0 74.0
```

In [34]:
```
IQR = q3-q1
```

In [35]:
```
lwr_bound = q1-(1.5*IQR)
upr_bound = q3+(1.5*IQR)
print(lwr_bound, upr_bound)
```

```
49.0 89.0
```

In [36]:
```
r_outliers = []
for i in sorted_rscore:
    if(i<lwr_bound or i>upr_bound):
        r_outliers.append(i)
print(r_outliers)
```

```
[95, 95, 96]
```

In [37]:
```
new_df = df1
for i in sample_outliers:
    new_df.drop(i,inplace=True)
new_df
```

Out[37]:

	math score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60	63	76	95	2021	3	female
1	75	70	64	85	2020	3	male
2	74	50	55	91	2020	3	male
3	68	76	78	97	2020	3	female
4	94	67	71	93	2020	3	male
6	61	78	92	94	2021	3	male
7	61	74	78	80	2021	2	male
8	64	76	79	76	2019	2	male
9	65	95	75	90	2020	3	female
10	66	76	67	100	2019	1	male
11	84	67	71	92	2020	3	male
12	69	66	70	56	2021	3	female
13	74	65	65	80	2021	2	male
14	74	63	72	96	2018	3	male
15	76	64	80	96	2020	3	male
16	60	64	54	91	2021	3	female
17	77	70	72	99	2020	3	male
18	67	95	64	87	2018	3	female
20	58	65	96	92	2019	3	female
21	68	63	62	94	2021	3	male
22	77	63	68	97	2021	3	female
23	80	64	86	85	2018	3	female
24	84	63	67	83	2018	1	male
25	68	67	73	88	2019	3	female
26	76	64	68	96	2021	3	female
27	92	96	61	83	2018	2	male
28	60	68	59	93	2020	3	male

In [38]:
```
df_stud = df1
ninetieth_percentile = py.percentile(df_stud['math score'], 90)
b = py.where(df_stud['math score']>ninetieth_percentile,ninetieth_percentile
print("New array:" ,b)
```

```
New array: [60. 75. 74. 68. 84. 61. 61. 64. 65. 66. 84. 69. 74. 74. 76. 6
0. 77. 67.
 58. 68. 77. 80. 84. 68. 76. 84. 60.]
```

In [39]:
```
df_stud.insert(1, "m score" , b, True)
df_stud
```

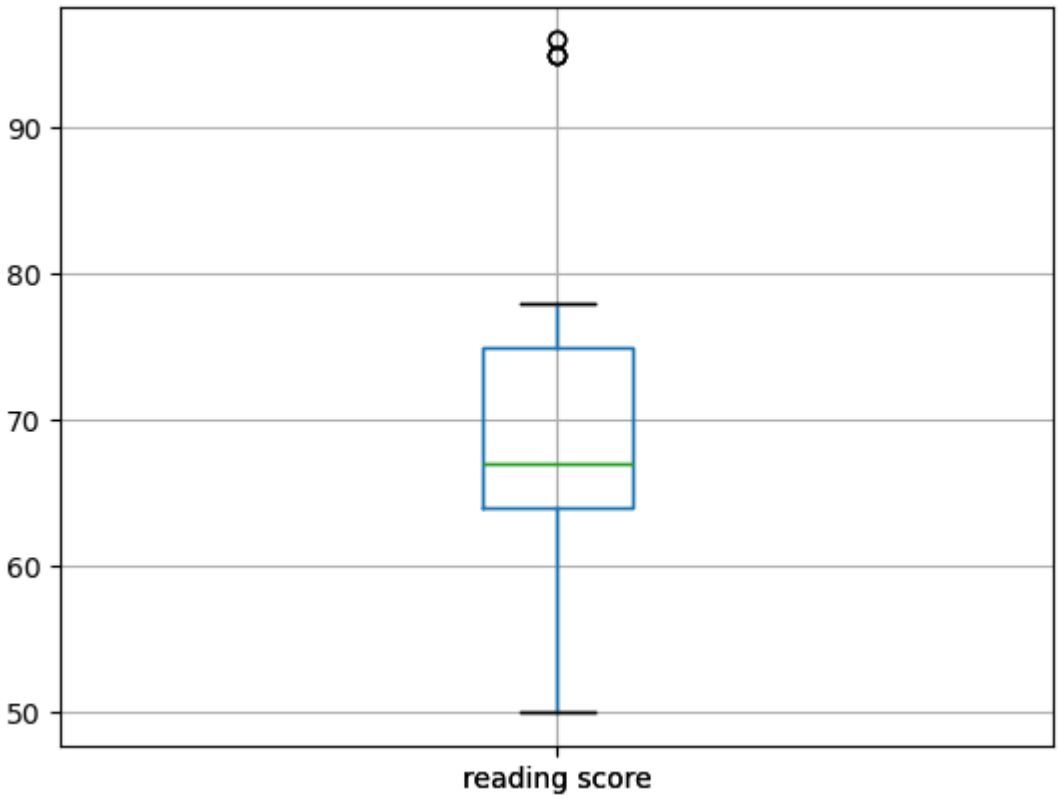
Out[39]:

	math score	m score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60	60.0	63	76	95	2021	3	female
1	75	75.0	70	64	85	2020	3	male
2	74	74.0	50	55	91	2020	3	male
3	68	68.0	76	78	97	2020	3	female
4	94	84.0	67	71	93	2020	3	male
6	61	61.0	78	92	94	2021	3	male
7	61	61.0	74	78	80	2021	2	male
8	64	64.0	76	79	76	2019	2	male
9	65	65.0	95	75	90	2020	3	female
10	66	66.0	76	67	100	2019	1	male
11	84	84.0	67	71	92	2020	3	male
12	69	69.0	66	70	56	2021	3	female
13	74	74.0	65	65	80	2021	2	male
14	74	74.0	63	72	96	2018	3	male
15	76	76.0	64	80	96	2020	3	male
16	60	60.0	64	54	91	2021	3	female
17	77	77.0	70	72	99	2020	3	male
18	67	67.0	95	64	87	2018	3	female
20	58	58.0	65	96	92	2019	3	female
21	68	68.0	63	62	94	2021	3	male
22	77	77.0	63	68	97	2021	3	female
23	80	80.0	64	86	85	2018	3	female
24	84	84.0	63	67	83	2018	1	male
25	68	68.0	67	73	88	2019	3	female
26	76	76.0	64	68	96	2021	3	female
27	92	84.0	96	61	83	2018	2	male
28	60	60.0	68	59	93	2020	3	male

In [41]:
```
col1 = ['reading score']
df1.boxplot(col1)
```

Out[41]: <Axes: >

In [42]:
```
plt.show()
```



In [43]:
```
median = py.median(sorted_rscore)
median
```

Out[43]:
```
66.0
```

In [44]:
```
refined_df = df1
refined_df['reading score'] = py.where(refined_df['reading score']>upr_bound
```

In [45]:
```
refined_df
```

Out[45]:

	math score	m score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60	60.0	63.0	76	95	2021	3	female
1	75	75.0	70.0	64	85	2020	3	male
2	74	74.0	50.0	55	91	2020	3	male
3	68	68.0	76.0	78	97	2020	3	female
4	94	84.0	67.0	71	93	2020	3	male
6	61	61.0	78.0	92	94	2021	3	male
7	61	61.0	74.0	78	80	2021	2	male
8	64	64.0	76.0	79	76	2019	2	male
9	65	65.0	66.0	75	90	2020	3	female
10	66	66.0	76.0	67	100	2019	1	male
11	84	84.0	67.0	71	92	2020	3	male
12	69	69.0	66.0	70	56	2021	3	female
13	74	74.0	65.0	65	80	2021	2	male
14	74	74.0	63.0	72	96	2018	3	male
15	76	76.0	64.0	80	96	2020	3	male
16	60	60.0	64.0	54	91	2021	3	female
17	77	77.0	70.0	72	99	2020	3	male
18	67	67.0	66.0	64	87	2018	3	female
20	58	58.0	65.0	96	92	2019	3	female
21	68	68.0	63.0	62	94	2021	3	male
22	77	77.0	63.0	68	97	2021	3	female
23	80	80.0	64.0	86	85	2018	3	female
24	84	84.0	63.0	67	83	2018	1	male
25	68	68.0	67.0	73	88	2019	3	female
26	76	76.0	64.0	68	96	2021	3	female
27	92	84.0	66.0	61	83	2018	2	male
28	60	60.0	68.0	59	93	2020	3	male

In [46]:
```
refined_df['reading score'] = py.where(refined_df['reading score']<lwr_bound
refined_df
```

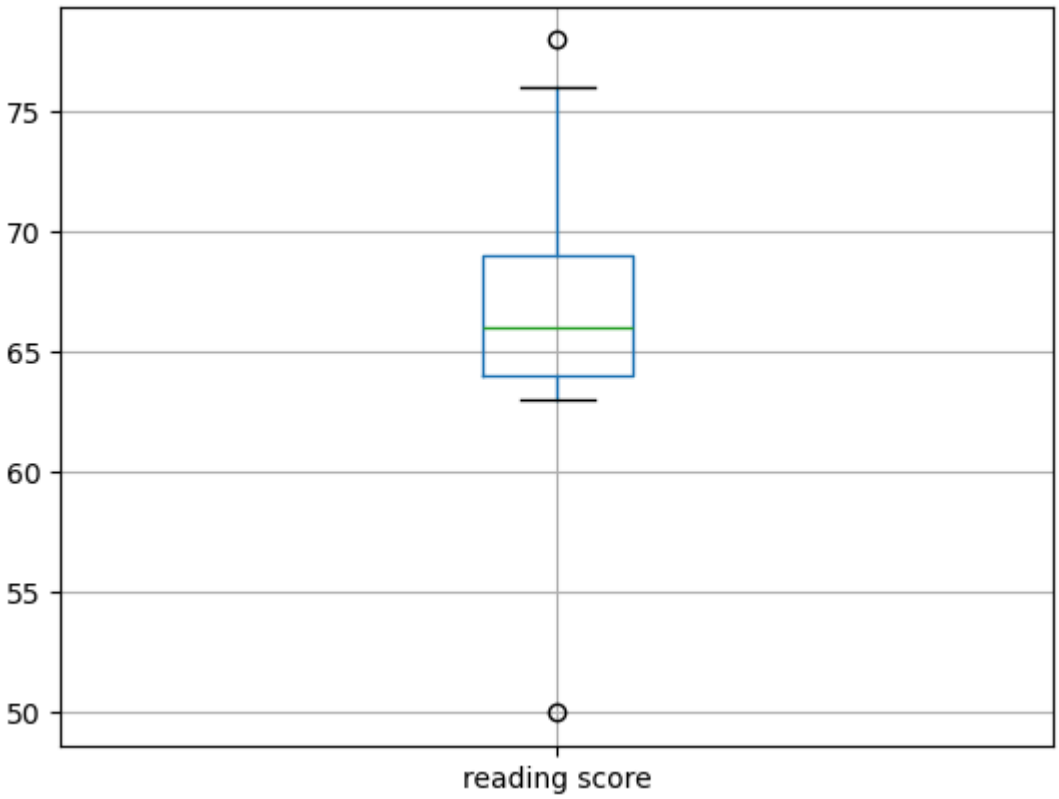
Out[46]:

	math score	m score	reading score	writing score	placememt score	club join year	placement offer count	gender
0	60	60.0	63.0	76	95	2021	3	female
1	75	75.0	70.0	64	85	2020	3	male
2	74	74.0	50.0	55	91	2020	3	male
3	68	68.0	76.0	78	97	2020	3	female
4	94	84.0	67.0	71	93	2020	3	male
6	61	61.0	78.0	92	94	2021	3	male
7	61	61.0	74.0	78	80	2021	2	male
8	64	64.0	76.0	79	76	2019	2	male
9	65	65.0	66.0	75	90	2020	3	female
10	66	66.0	76.0	67	100	2019	1	male
11	84	84.0	67.0	71	92	2020	3	male
12	69	69.0	66.0	70	56	2021	3	female
13	74	74.0	65.0	65	80	2021	2	male
14	74	74.0	63.0	72	96	2018	3	male
15	76	76.0	64.0	80	96	2020	3	male
16	60	60.0	64.0	54	91	2021	3	female
17	77	77.0	70.0	72	99	2020	3	male
18	67	67.0	66.0	64	87	2018	3	female
20	58	58.0	65.0	96	92	2019	3	female
21	68	68.0	63.0	62	94	2021	3	male
22	77	77.0	63.0	68	97	2021	3	female
23	80	80.0	64.0	86	85	2018	3	female
24	84	84.0	63.0	67	83	2018	1	male
25	68	68.0	67.0	73	88	2019	3	female
26	76	76.0	64.0	68	96	2021	3	female
27	92	84.0	66.0	61	83	2018	2	male
28	60	60.0	68.0	59	93	2020	3	male

In [47]:
```
col2 = ['reading score']
refined_df.boxplot(col2)
```

Out[47]: <Axes: >

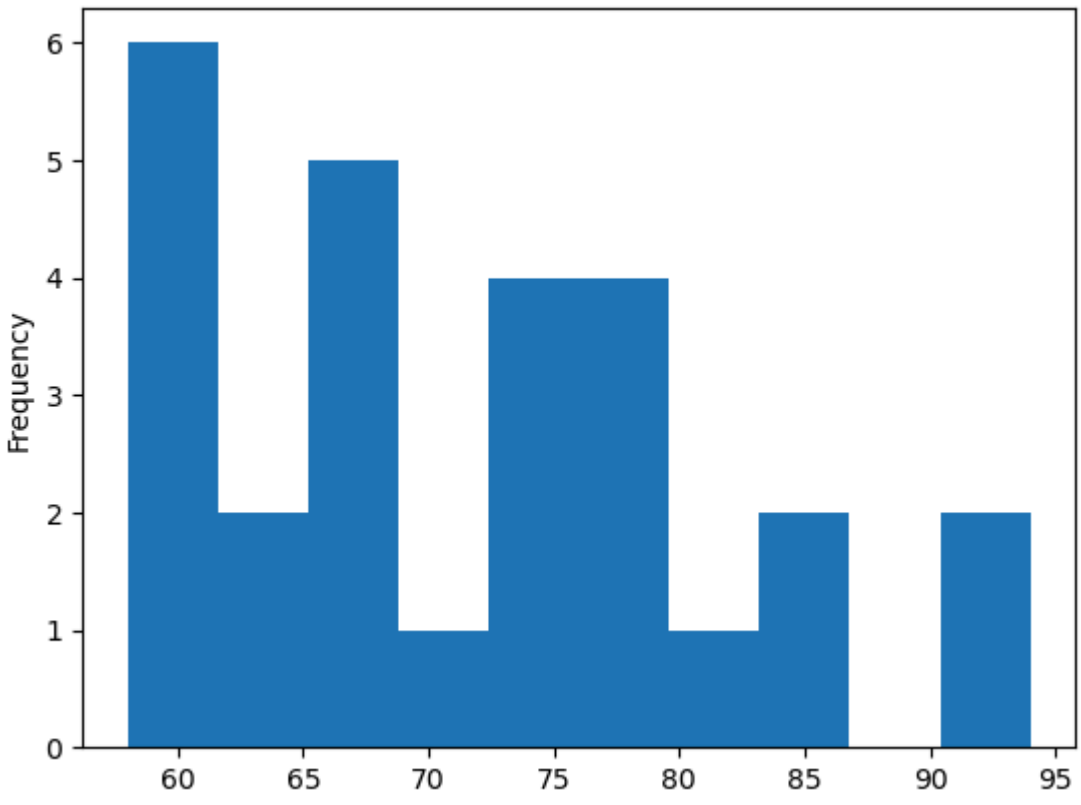
In [48]:
```
plt.show()
```



In [50]:
```
new_df['math score'].plot(kind = 'hist')
```

Out[50]:
```
<Axes: ylabel='Frequency'>
```

In [51]:
```
plt.show()
```

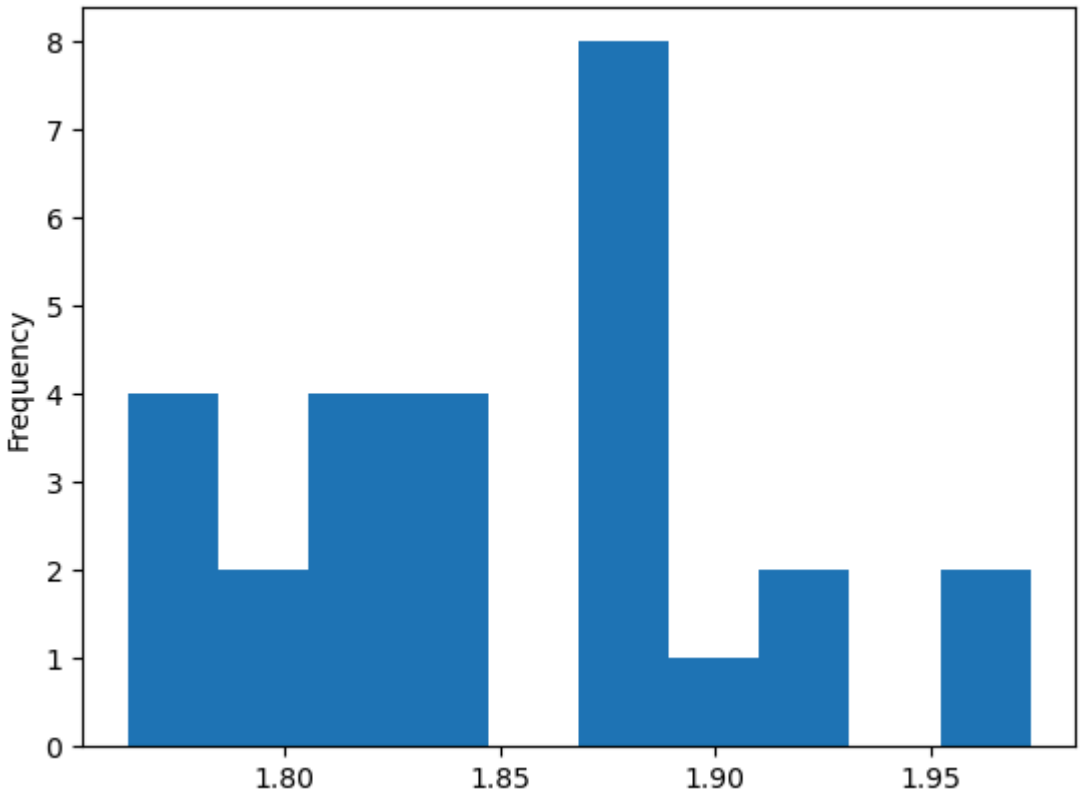


In [52]:
```
df1['log_math'] = py.log10(df1['math score'])
```

In [53]:
```
df1['log_math'].plot(kind = 'hist')
```

Out[53]:
```
<Axes: ylabel='Frequency'>
```

In [54]:
```
plt.show()
```



In [ ]:
```

```